(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平7－78155

(43)公開日　平成7年(1995)3月20日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ	技術表示箇所
Ｇ０６Ｆ　17/22				
Ｇ０６Ｋ　9/72		8623－5Ｌ		
		7315－5Ｌ	Ｇ０６Ｆ　15/20　５２８　Ｆ	

審査請求　未請求　請求項の数4　ＦＤ　（全７頁）

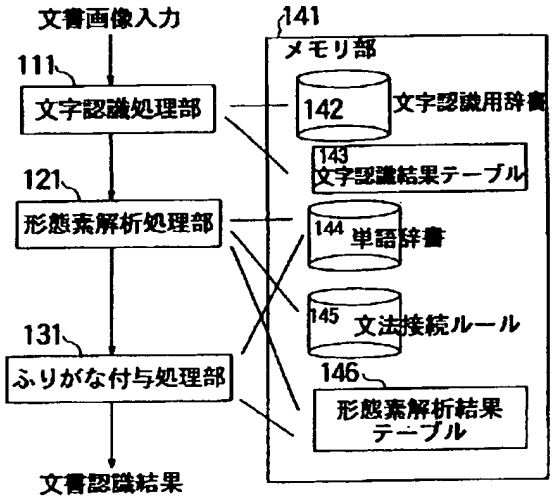
(21)出願番号　特願平5－172721

(22)出願日　平成5年(1993)6月18日

(71)出願人　000102728
エヌ・ティ・ティ・データ通信株式会社
東京都江東区豊洲三丁目3番3号
(72)発明者　原　恵子
東京都江東区豊洲三丁目3番3号　エヌ・ティ・ティ・データ通信株式会社内
(72)発明者　吉野　順
東京都江東区豊洲三丁目3番3号　エヌ・ティ・ティ・データ通信株式会社内
(72)発明者　岩根　和巳
東京都江東区豊洲三丁目3番3号　エヌ・ティ・ティ・データ通信株式会社内
(74)代理人　弁理士　上村　輝之

(54)【発明の名称】　文書認識装置

(57)【要約】
【目的】　認識文書について、１度だけの形態素解析によって表記と読みとの双方を出力できる文書認識装置を提供する。
【構成】　文字認識処理部（１１１）が、入力文書の文字列画像から文字を切出して、文字認識を行なって複数個の認識候補文字列を出力する。次に、形態素解析処理部（１２１）が、認識候補文字列から単語を抽出し、各単語の接続検定を行い、接続可能な単語を繋げていくことにより、入力文書の正しい表記を決定する。その際、形態素解析処理部（１２１）は更に、入力文書の各単語に単語番号を付与して、正しい表記と共に出力する。次に、ふりがな付与部（１３１）が、形態素解析部から得た文の表記に、各単語の単語番号に基づいて、予め用意したふりがなを自動的に付与する。



1

【特許請求の範囲】

【請求項1】　入力文書の文字列画像について文字認識を行なって、認識候補文字列を出力する文字認識処理手段と、

前記認識候補文字列について形態素解析を行なって、正解文の表記と、正解文を構成する各単語の単語番号とを出力する形態素解析処理手段と、

前記形態素解析手段より出力した単語番号に基づいて、前記正解文に自動的にふりがなを付与する自動ふりがな付与手段とを備えることを特徴とする文書認識装置。

【請求項2】　請求項1記載の装置において、前記形態素解析に必要な種々の単語の表記と単語番号、及び前記ふりがな付与処理に必要な単語番号に対応する読みを格納した単語辞書を備え、

前記単語辞書では、表記が同一で読みが異なる単語が別単語として登録されていることを特徴とする文書認識装置。

【請求項3】　請求項2記載の装置において、前記単語辞書には、漢字の単語の読みのみが登録されていることを特徴とする文書認識装置。

【請求項4】　請求項1乃至3のいずれか記載の装置において、前記形態素解析手段が、前記正解文の表記と共に、この正解文の各単語を区切るマーキングと各単語の単語番号とを出力し、

前記ふりがな付与手段が、前記正解文の各単語の単語番号に対応する読みを前記単語辞書から取得することにより、前記正解文にふりがなを付与することを特徴とする文書認識装置。

【発明の詳細な説明】

【0001】

【産業上の利用分野】本発明は文書認識装置に関するものであり、特に、出力文書に自動的にふりがなを付与することができる文書認識装置に関するものである。

【0002】

【従来の技術】従来、文書認識装置は、文字認識部と形態素解析部とを備える。文字認識部では入力した文書の文字列画像から文字を切出して文字認識を行なって、複数個の認識候補文字を出力する。形態素解析部は、文字認識部で認識した認識候補文字を単語辞書と照合して、照合単語間の接続検定を行ない、接続可能な単語の組合わせから出力文書を作成する。

【0003】従来の文書認識装置における形態素解析部は、認識候補文字の表記のみを用い、したがって、出力結果も認識した文字の表記のみであり、これらの文字の読みであるふりがなまでは問題にしていない。

【0004】

【発明が解決しようとする課題】このように、従来の文書認識装置の出力結果には、文書の読み、すなわちふりがなが付与されていないため、例えば、文字認識装置を文字読上げ装置と接続して、盲人向けの文章朗読を自動

2

化する場合等、出力結果を音声化する必要がある場合に、文字認識装置の出力結果について再度形態素解析を行なって出力文字の読みを得る必要がる。つまり、従来は、文書認識装置で認識した文書を音声化するにあたって、認識候補文字から認識結果を得るためと、この認識結果の読みを得るためとに、2度形態素解析を行なう必要がある。このため装置の構成が複雑になり、処理時間が長くなるという問題がある。従って、本発明の目的は、1度だけの形態素解析によって認識文書について、表記と読みとの双方を出力できる文書認識装置を提供する。

【0005】

【課題を解決するための手段】上記目的を達成するために、本発明は、入力文書の文字列画像について文字認識を行なって、認識候補文字列を出力する文字認識処理手段と、この認識候補文字列について形態素解析を行なって、正解文の表記と、正解文を構成する各単語の単語番号とを出力する形態素解析処理手段と、形態素解析手段より出力した単語番号に基づいて、前記正解文に自動的にふりがなを付与する自動ふりがな付与手段とを備えることを特徴とする文書認識装置。

【0006】

【作用】入力文書について、まず文字認識を行なって、認識候補文字列を得る。次に、この認識候補文字列に対して形態素解析を行なって、正解文の表記を決定する。形態素解析では一般に、単語辞書を参照して、認識候補文字列から各単語を抽出し、単語間の接続検定を行ない、接続可能な単語を繋げていくことにより正解文の表記を決定する。

【0007】この形態素解析において、上記表記決定処理に加え、各単語にその単語番号ををを付す処理が行なわれる。これにより、形態素解析から得られた正解文は、その表記に各単語の単語番号が付加されたものとなる。この正解文に対し、各単語番号に基づき各単語の読み（ふりがな）が付与される。この結果、読みが追加された正解文が出力される。

【0008】このようにして、従来、文書を認識するためと、その読みを得るために2度必要であった形態素解析処理を一度行なうだけで、文書の表記とその読みとを得ることができるため、処理の高速化を図ることができる。

【0009】

【実施例】以下に図面を用いて本発明の文書認識装置の実施例を説明する。

【0010】図1は、本発明の文書認識装置の一実施例構成を示すブロック図である。

【0011】図1に示すとおり、文書認識装置は、入力文書画像の文字認識を行なう文字認識処理部111、文字認識処理部111で認識した認識文字について形態素解析を行なう形態素解析処理部121、形態素解析結果

にふりがなを付与するふりがな付与処理部１３１、及びメモリ部１４１とで構成されている。

【００１２】メモリ部１４１は、文字の画像特徴量を記録した文字認識用辞書１４２、文字認識の結果を格納する文字認識結果テーブル１４３、単語の表記、読み、文法的な接続情報を記録した単語辞書１４４、単語間接属のルールを記述した文法接続ルール１４５、及び形態素解析結果を格納する形態素解析結果テーブル１４６とから構成されている。

【００１３】図1に示す文書認識装置における文書認識処理を、以下に説明する。

【００１４】まず、文字認識処理部１１１に文書画像を入力し、入力文書から入力文字列画像の文字切出しを行なって、文字認識用辞書１４２を用いて文字認識を行なう。この認識結果（複数個の認識候補文字列）は文字認識結果テーブル１４３に格納しておく。

【００１５】ついで、形態素解析処理部１２１にて、文字認識結果テーブル１４３に格納した認識候補文字列について形態素解析を行なって、認識候補文字列から正解文字を選択した文、あるいは認識候補文字列に正解文字が含まれていない場合はこれを訂正した文を作成する。この形態素解析では、認識候補文字列を単語辞書１４４と照合して単語を抽出し、更に、この抽出した単語について、文法接続ルール１４５による単語間の接続検定を行ない、接続可能な単語を繋げていくことにより、出力文を作成する。なお、単語辞書１４４に格納されている各単語には、一意に単語番号が与えられており、形態素解析処理部１２１は、作成した出力文について、その文の表記と共に、この文を構成する単語間を区切るマーキングと、各単語の単語番号とを付加した情報を、形態素解析結果テーブル１４６に格納する。

【００１６】図2は、単語辞書１４４の構成を示す図である。単語辞書１４４は、各単語の表記品詞、接続可能な品詞リスト等の形態素解析用の単語情報を単語番号順に格納した形態素解析用テーブル１４４ａ、ふりがなテーブル１４４ｃへのポインタを単語番号順に格納した単語情報テーブル１４４ｂ、及び、単語番号順にふりがなのみを格納したふりがなテーブル１４４ｃとから構成されている。なお、この単語辞書１４４では、表記が同一で読みが異なる単語については別単語として登録されている。

【００１７】ふりがな付与処理部１３１において、形態素解析結果テーブル１４６に格納した文について、この文を構成する各単語に付された単語番号に基づいて単語辞書１４４から単語の読みを得て、出力文にふりがなを付与する。なお、単語辞書１４４でふりがなを別テーブルに格納するのは、ふりがなの長さが単語によって異なるためである。

【００１８】ふりがな付与処理部１３１では、出力文を構成する各単語の単語番号を取得し、この単語番号に対応する単語について、ふりがなテーブル１４４ｃを参照してふりがなを取得する。ついで、形態素解析結果テーブル１４６から出力文書の表記を得て、先に取得したふりがなと共に文書認識結果として出力する。

【００１９】次に、図１に示す文書認識装置における処理の具体例を図３及び図４を用いて説明する。

【００２０】この例では手書きされた「重箱読みと湯桶読みがある」が入力文書であるものとする。

【００２１】まず、文字認識処理部１１１において、入力した画像に対して文字の切出しを行ない、文字認識用辞書１４２を用いて文字認識を行なう。図3に、この結果得られたＮ個の認識候補文字列を示す。このようにして得た認識候補文字列は、文字認識結果テーブル１４３に格納される。

【００２２】次に、形態素解析処理部１２１においてこの認識候補文字列に対して形態素解析を行なって、正解文字を選択し、或いは正解文字が含まれていない場合はこの訂正を行なう。即ち、認識候補文字列と単語辞書１４４の形態素解析用テーブル１４４ａに記録した品詞情報、表記等を照合して単語を抽出し、文法接続ルール１４５を用いて抽出した単語間の接続検定を行なう。そして、接続可能である単語組を繋げていくことによって、出力文を構成してその表記を形態素解析結果格納テーブル１４６に格納する。その際、形態素解析用テーブル１４４ａに記録されている単語情報には、単語番号が付与されているので、これを参照して出力文を構成する各単語を区切るマーキングと各単語の単語番号とを、出力文の表記と共に形態素解析結果テーブル１４６に格納するようにする。この例では、入力文通り解析され、「重箱読み｜と｜湯桶読み｜が｜ある」と出力されたものとする。

【００２３】形態素解析の手法としては、右方向最長一致法を用いる手法、接続表を用いる手法等があるが、どの手法を用いてもよい。

【００２４】次に、ふりがな付与処理部１３１において、上記形態素解析結果に対してふりがなの付与を行なう。即ち、上記の形態素解析処理部１２１の出力文書を構成する単語に付されている単語番号に基づいて当該単語の読みを得るようにする。即ち、ふりがなテーブル１４４ｃから各単語の単語番号に対応するふりがなを取得して、認識文書の表記と共に、この表記の読み（ふりがな）を付して、ふりがな付きの文書として出力する。

【００２５】なお、単語辞書１４４に、漢字部分の読みのみを登録するようにしておくことによって、出力文書の漢字部分についてのみふりがなを付与することが可能である。

【００２６】以上、本発明の好適な一実施例を説明したが、本発明はこの実施例にのみ限定されるのではなく、種々の異なる態様で実施することが可能である。

【００２７】

5

【発明の効果】上記に詳細に説明したとおり、本発明の文書認識装置では、文書の表記を決定するための形態素解析において、文書を構成する各単語と、予め用意した各単語の読みとの対応付けを行なうようにしているので、この文書を音声化して出力する際等に、従来、2度行なう必要のあった形態素解析処理を一度で済ませることができ、よって処理の高速化を図ることが可能となる。

【図面の簡単な説明】

【図1】　本発明の文書認識装置の構成を示すブロック図である。

【図2】　本発明の装置に用いる単語辞書の構成を示す図である。

【図2】　図1に示す装置における処理の具体例を示す

6

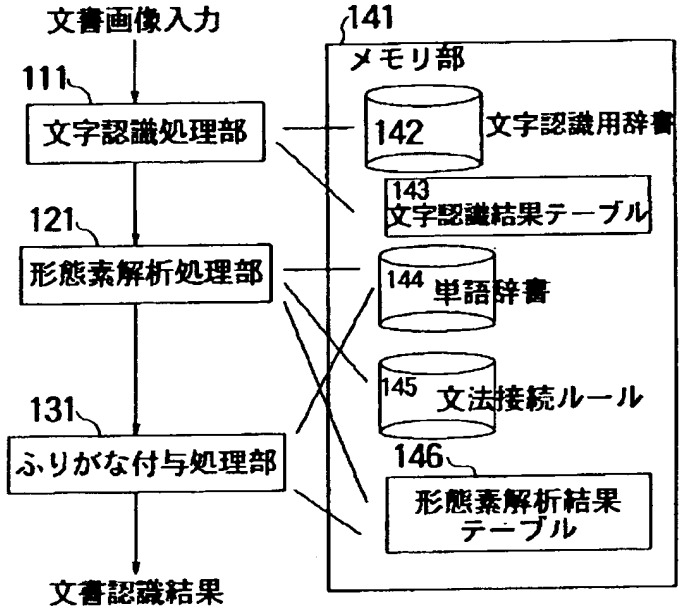
図である。

【図3】　入力データの文書例とその文字認識結果の一例を示す図である。

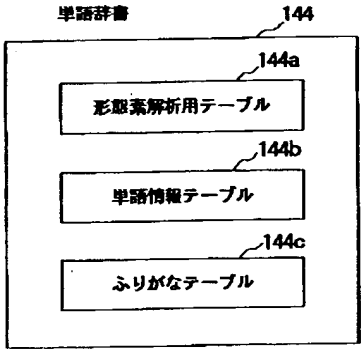
【符号の説明】

111　文字認識処理部
121　形態素解析処理部
131　ふりがな付与処理部
141　メモリ部
142　文字認識用辞書
143　文字認識結果テーブル
144　単語辞書
145　文法接続ルール
146　形態素解析結果テーブル

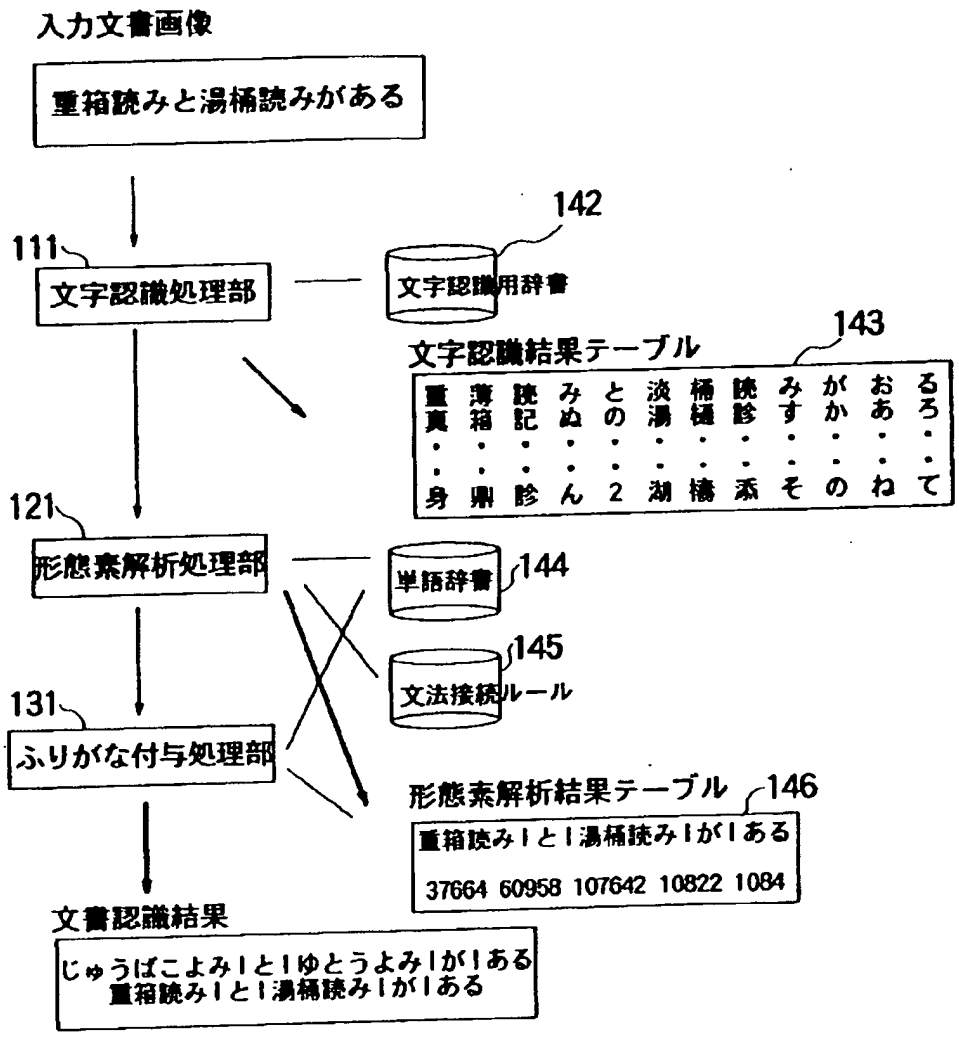
【図1】



【図2】



【図3】


入力文書画像
重箱読みと湯桶読みがある
111
文字認識処理部
142
文字認識用辞書
文字認識結果テーブル
143
重真・・身
薄箱・・鼎
読記・・診
みぬ・・ん
との・・2
淡湯・・湖
桶樋・・橋
読診・・添
みす・・そ
がか・・の
おあ・・ね
るろ・・て
121
形態素解析処理部
単語辞書
144
145
文法接続ルール
131
ふりがな付与処理部
形態素解析結果テーブル
146
重箱読み|と|湯桶読み|が|ある
37664 60958 107642 10822 1084
文書認識結果
じゅうばこよみ|と|ゆとうよみ|が|ある
重箱読み|と|湯桶読み|が|ある

【図4】

1文字目 2文字目 3文字目 4文字目

入力文	重	箱	読	み	と	湯	桶	読	み	が	あ	る
第1位	重	薄	読	み	と	淡	桶	読	み	が	お	る
第2位	真	箱	記	ぬ	の	湯	樋	診	す	か	あ	ろ
第3位	見	箸	熟	す	せ	添	通	誠	ぬ	さ	の	ぬ
	・	・	・	・	・	・	・	・	・	・	・	・
第N位	身	鼎	診	ん	2	湖	橋	添	そ	の	ね	て

【手続補正書】
【提出日】平成6年3月16日
【手続補正1】
【補正対象書類名】明細書
【補正対象項目名】図面の簡単な説明
【補正方法】変更
【補正内容】
【図面の簡単な説明】
【図1】　本発明の文書認識装置の構成を示すブロック図である。
【図2】　本発明の装置に用いる単語辞書の構成を示す図である。
【図3】　図1に示す装置における処理の具体例を示す図である。
【図4】　入力データの文書例とその文字認識結果の一例を示す図である。
【符号の説明】
111　文字認識処理部
121　形態素解析処理部
131　ふりがな付与処理部
141　メモリ部
142　文字認識用辞書
143　文字認識結果テーブル
144　単語辞書
145　文法接続ルール
146　形態素解析結果テーブル